

International Journal of Population Data Science

Journal Website: www.ijpds.org



Probabilistic Record Linkage of 90 Million Cervical Cancer-Related Records in South Africa

Peace Ayeni¹, Carole Metekoua², Matshediso Mohlala³, Jenny Matthew³, Sizeka Mashele⁴, Judith Mwansa³, Tafadzwa Dhokotera¹, Meredith Apelberg⁵, Aimee Kreimer⁵, and Eliane Rohner⁶

¹Swiss Tropical and Public Health Institute (Swiss TPH), Allschwil, Switzerland; University of Basel, Basel, Switzerland

²National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa; University of Bern, Bern, Switzerland

³National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa

⁴National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa; Swiss Tropical and Public Health Institute (Swiss TPH), Allschwil, Switzerland; University of Basel, Basel, Switzerland

⁵National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

⁶Institute of Social and Preventive Medicine (ISPM), Bern, Switzerland; University of Bern, Bern, Switzerland

We used probabilistic record linkage (PRL) to create the population-based, nation-wide South African Cervical Cancer Screening (SACCS) cohort and evaluated linkage quality. We linked cervical cytology screening, histology, HIV, and cancer records from the National Health Laboratory Service and the National Cancer Registry of South Africa. Linkage variables included given names, surname and date or year of birth. We used G-Link, a PRL software based on the Fellegi-Sunter RL algorithm, with exact matching preceding PRL. Cytology and HIV datasets were appended and deduplicated. To reduce computation costs, we trained the model on 10% of records with national IDs (reference standard) and tested on all ID-bearing records to select the best threshold, which we then applied to deduplicate all records. After generating unique person IDs, cytology records were linked to cancer records

using a two-file method. We used the same two-file linkage steps for cytology-histology linkage, with thresholds chosen by optimal precision, recall and F-measure. We extracted 12.4M cervical screening cytology, 1.3M histology, 669k cancer records, all from 2005-2023 and 74.2M HIV records (2004-2021). Deduplicated cytology-HIV linkage model achieved 99% precision, 97% recall and 98% F-measure. Cytology-cancer linkage reached 98% precision, 90% recall and 94% F-measure. Cytology-histology preliminary linkage results were 96% precision, 89% recall and 93% F-measure. The final cohort included 8,787,734 unique women with at least one cytology record, 246,091 histology reports, and 171,669 cancer diagnoses. Robust linkage performance establishes SACCS as a reliable resource for monitoring progress towards cervical cancer elimination in South Africa.

